ETHICS OF ARTIFICIAL INTELLIGENCE

A SOCIAL CATASTROPHE?

Arvind Krishna Menon

SCMS School of Technology and Management

Kochi, India

arvindmenon0103@gmail.com

**ABSTRACT—the paper mainly talks about how an artificial system can accommodate a relationship with human beings without disrupting the ecosystem. Some ethical issues like machines harming humans, and other morally relevant beings are taken into concern. The paper is divided into 5 sections. The first section deals with issues regarding its creation and development. The second section describes the challenges when AI approach human behaviour. The third section highlights how and in what scenario AI machines have a moral status of their own. The final section deals with the issues of creating AI more intelligent than human and ensure the technology is not used for any ill purposes.**

**Index terms- Artificial intelligence, ethics, humans**

1. INTRODUCTION

A

rtificial intelligence has come a long way since its inception. It began with simple philosophies and speculations of how we could develop something inanimate, which performs human-like tasks as a symbolic system. But it wasn’t until recent times the word “Artificial Intelligence” was officially coined. The basis of artificial intelligence can be defined in a rather simple way- to mimic and perform tasks, both simple and complex as the industry aims to include learning, reasoning and perception. One of the key component which constitutes AI machine are algorithms. Simple processes require simple algorithms and vice versa. The evolution of AI is beneficial to many countries as it makes daily tasks much easier. Some of the better known applications include healthcare, finance and trading, quantum mechanics and so on.

Most researchers believe that a super intelligent AI system will be unable to feel human like emotions like love or hate. So there is no reason to expect AI to intentionally become harmful to human beings. There are two likely scenarios which could take place,

* 1. The AI is programmed to do something destructive.
  2. It is programmed to do something beneficial, but it develops a potentially destructive path to achieve it.

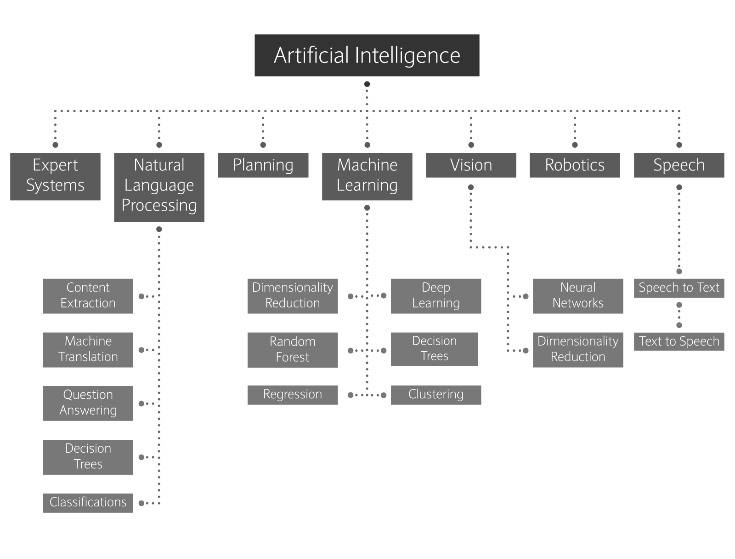
So the accusation of AI themselves becoming malevolent to humans is a bit farfetched. An intelligent system will be good at accomplishing goals since that is what it is programmed to do. If those goals aren’t aligned with ours, then we have a problem. Because AI has the potential to become much more intelligent than humans and there is no sure-fire way to know how it will behave. We can’t use technological development because intelligence cannot be “programmed”. It evolves from experience, and its ability to learn from everything.

An example of situation getting worse is in the case of autonomous weapons. Autonomous weapons are AI systems programmed to kill. In the hands of the wrong person, autonomous weapons could lead to an AI war, and mass casualties, potentially even the end of humanity. Such weapons may be designed to be extremely difficult to “turn off”, and humans could plausibly, rapidly lose control. This risk is prevalent even with narrow AI, but grows exponentially as autonomy increases.

We have 2 types of AI- strong and weak AI. The distinction between the both is quite simple. A weak AI will always have a programmed response, which means it will always have a definite response the programmer has embedded while, the strong AI is much more complex. In one way we can say it responds the way, a human does i.e. it doesn’t have a set answer to your keywords. The system will provide one but can’t be sure of the result. There is also one other type of AI known as super AI which can be considered as the best of the 3. It surpasses human intelligence and ability. In a nutshell it is best at everything. So far, we’ve only achieved the first of the three types of AI — weak AI. As research continues, it’s reasonable to strive for strong AI.

The picture below depicts the sub-divisions of AI:

Each of them serves different purpose in terms of applications but all of them use algorithms (and statistics) to replicate or stimulate human intelligence.



1. ETHICS IN MACHINE LEARNING AND THE NEAR FUTURE OF AI

Suppose we create a system which works for a bank. It helps us select candidates who have applied for loan. Now a typical AI system would select a person irrespective of factors like gender, ethnicity, class etc. But what if the machine did racially discriminate among candidates? The AI strictly evaluates on algorithms and information provided to it. After analysis it showed that black people are less prone to pay back and hence why the future candidate was denied of the loan. This basically implies that the algorithm accepts white applicants and rejects black applicants.

The answer to why this happens may not be that easy. There are two ways by which we can program an algorithm. The first is based on a complicate neural network whereas the second uses Bayesian network and decision trees. The latter is direct in its approach and finding the issue can be simple for this model. But the first one analyse data more “human-like” using genetic algorithm or directed evolution. Just like we can’t predict how a human would react to a situation, we won’t be able to understand why the AI made that decision.

AI algorithms play an increasingly large role in modern society. It will become increasingly important to develop AI algorithms that are not just powerful and scalable, but also transparent to inspection. Some challenges of machine ethics are much like many other challenges involved in designing machines.

Transparency is not the only desirable feature of AI. It is also important that AI algorithms taking over social functions be *predictable to those they govern*. It will also become increasingly important that AI algorithms be robust against manipulation.

Another concern is who will be held responsible for getting something done or when the system fails at something. Modern bureaucrats often take refuge in established procedures that distribute responsibility so widely that no one person can be identified to blame. Even if an AI system is designed with a user override, one must consider the career incentive of a bureaucrat who will be personally blamed if the override goes wrong.

Responsibility, transparency, auditability, incorruptibility, predictability, and a tendency to not make innocent victims scream with helpless frustration: all are criteria that must be considered in an algorithm intended to replace human judgment of social functions.

1. ARTIFICIAL GENERAL INTELLIGENCE(AGI)

An Artificial General Intelligence would be a machine capable of understanding the world as well as any human, and with the same capacity to learn how to carry out a huge range of tasks. There is a universal agreement among modern AI professionals that Artificial Intelligence falls short of human capabilities in some sense, even though AI algorithms have beaten humans in many specific domains such as chess—chess was considered the epitome of intelligence until Deep Blue won the world championship from Kasparov—but even these researchers agree that something important is missing from modern AIs.

It is debatable whether human intelligence is truly general—we are certainly better at some cognitive tasks than others but human intelligence is surely significantly more generally applicable than non-hominid intelligence.

It is a qualitatively different class of problem to handle an AGI operating across many novel contexts that cannot be predicted in advance. Even task-specific AI algorithms give hard times to predict. This can be seen in the case of Deep Blue and how it won against grandmaster Garry Kasparov.

In 1997 deep blue defeated the reigning chess champion Garry Kasparov. How the AI actually managed to that is quite interesting. Programming it the normal way by inserting every rules and steps, to ensure complete victory is a daunting and near to impossible task. Also the programmers had little to no knowledge of the game to play professionally. So inserted a “wining region” as defined by the chess rules. This predicted distant consequences. Though it proved accurate, did not allow the programmers to envision the local behaviour of Deep Blue. There is particular answer to why the AI made a particular move.

The discipline of AI ethics, especially as applied to AGI, is likely to differ fundamentally from the ethical discipline of non-cognitive technologies. This is explained as follows:

* Verifying the safety of the system becomes a greater challenge because we must verify what the system is trying to do, rather than being able to verify the system’s safe behaviour in all operating contexts;
* Ethical cognition itself must be taken as a subject matter of engineering;
* The local, specific behaviour of the AI may not be predictable apart from its safety, even if the programmers do everything right.

1. MACHINES WITH MORAL STATUS

Technology has come a long way and innovations are inevitable. There will come a time when we will have to evaluate a machine based on its moral ethics or the types of ethics it should have in the first place. How should humans treat them and should they have any different ethics than the way we treat humans. Francis Kamm proposed the following definition of moral status, which will serve for our purposes:

*X has moral status = because X counts morally in its own right, it is permissible/impermissible to do things to it for its own sake*.

A rock has no moral status: we may crush it, pulverize it, or subject it to any treatment we like without any concern for the rock itself. A human person, on the other hand, must be treated not only as a means but also as an end. Exactly what it means to treat a person as an end is something about which different ethical theories disagree; but it certainly involves taking his/her legitimate interests into account and accepting moral behaviour when we interact with them- such as prohibition against murdering her, stealing from her, or doing a variety of other things to her or her property without her consent.

While it is fairly consensual that present-day AI systems lack moral status, it is unclear exactly what attributes ground moral status. Two criteria are commonly proposed as being importantly linked to moral status: sentience and sapience. These may be characterized roughly as follows:

* **Sentience**: the capacity for phenomenal experience or qualia, such as the capacity to feel pain and suffer.
* **Sapience**: a set of capacities associated with higher intelligence, such as self- awareness and being a reason-responsive agent.

One common view is that many animals have qualia and therefore have some moral status, but that only human beings have sapience, which gives them a higher moral status than non-human animals. This suggests that an AI system will have some moral status if it has the capacity for qualia, such as an ability to feel pain. If in addition to sentience, an AI system also has sapience of a kind similar to that of a normal human adult, then it would have full moral status, equivalent to that of human beings.

We must consider how these novel properties would affect the moral status of artificial minds and what it would mean to respect the moral status of such exotic minds.

1. SUPERINTELLIGENCE

Superintelligence is an intelligence system that rapidly increases its intelligence in a short time, specifically, to surpass the cognitive capability of the average human being. Superintelligence has been used in science fiction, and in discussions around artificial intelligence, to understand some of the consequences of a quickly evolving intelligence model in IT. Superintelligence is tied to the idea of a “singularity”, which is based on the idea that a catalyst or trigger would cause rapid change beyond what humans can anticipate.

Superintelligence may also be achievable by increasing processing speed. The fastest observed neurons fire 1000 times per second. It should be physically possible to build a brain which computes a million times as fast as a human brain, without shrinking its size or rewriting its software.

Yudkowsky lists three families of metaphors for visualizing the capability of a smarter-than-human AI:

* Metaphors inspired by differences of individual intelligence between humans.
* Metaphors inspired by knowledge differences between past and present human civilizations.
* Metaphors inspired by differences of brain architecture between humans and other biological organisms.

Even if we restrict ourselves to historical metaphors, it becomes clear that superhuman intelligence presents ethical challenges that are quite literally unprecedented. At this point the stakes are no longer on an individual scale proved, but on a global or cosmic scale.

Superintelligence is one of several “existential risks” we should really look put for. It is a risk “where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential.” Conversely, a positive outcome for superintelligence could preserve Earth-originating intelligent life and fulfil its potential.

One strong piece of advice that emerges from considering our situation as analogous to that of Archimedes is that we should not try to invent a “super” version of what our own civilization considers to be ethics—this is not the strategy we would have wanted Archimedes to follow. The questions we should be considering is how an AI programmed by Archimedes, with no more moral expertise than Archimedes, could recognize our own civilization’s ethics as moral progress as opposed to mere moral instability. This would require that we begin to comprehend the structure of ethical questions in the way that we have already comprehended the structure of chess.

If we are serious about developing advanced AI, this is a challenge that we must meet. If machines are to be placed in a position of being stronger, faster, more trusted, or smarter than humans, then the discipline of machine ethics must commit itself to seeking human niceness.

1. CONCLUSION

The current AI offers us a few ethical issues that are not already present in the design of cars or power plants. The approach of AI algorithms toward more humanlike thought can lead to predictable complications. General AI algorithms may no longer execute in predictable contexts, requiring new kinds of safety assurance and the engineering of artificial ethical considerations. AIs with advanced mental states will have moral status, and some may count as persons though it will be much unlike the sort that exist now. It could even be governed by different rules. And finally, the prospect of AIs with superhuman intelligence and superhuman abilities presents us with the extraordinary challenge of stating an algorithm that outputs super ethical behaviour. These challenges may seem visionary, but it seems predictable that we will encounter them; and they are not devoid of suggestions for present-day research directions.

1. REFERENCES
2. https://www.machinedesign.com/markets/robotics/article/21835139/whats-the-difference-between-weak-and-strong-ai
3. https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/
4. https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp#:~:text=Artificial%20intelligence%20(AI)%20refers%20to,as%20learning%20and%20problem%2Dsolving.
5. https://www.thinkautomation.com/bots-and-ai/types-of-ai-distinguishing-between-weak-strong-and-super-ai/
6. https://www.drishtiias.com/images/uploads/1579352054\_image5.png
7. https://www.raconteur.net/business-innovation/the-rise-of-artificial-intelligence-in-6-charts
8. https://www.techopedia.com/definition/32855/superintelligence
9. The Ethics Of Artificial Intelligence By Nick Bostrom And Eliezer Yudkowsky